Informatics in Education, 2017, Vol. 16, No. 2, 225–256

DOI: 10.15388/infedu.2017.12

Automatic Content Recommendation and Aggregation According to SCORM

Daniel Eugênio NEVES, Wladmir Cardoso BRANDÃO, Lucila ISHITANI

Institute of Exact Sciences and Informatics, Pontifical Catholic University of Minas Gerais Minas Gerais, Brazil
e-mail: danieleugenio.neves@gmail.com, wladmir@pucminas.br, lucila@pucminas.br

Received: February 2017

Abstract. Although widely used, the SCORM metadata model for content aggregation is difficult to be used by educators, content developers and instructional designers. Particularly, the identification of contents related with each other, in large repositories, and their aggregation using metadata as defined in SCORM, has been demanding efforts of computer science researchers in pursuit of the automation of this process. Previous approaches have extended or altered the metadata defined by SCORM standard. In this paper, we present experimental results on our proposed methodology which employs ontologies, automatic annotation of metadata, information retrieval and text mining to recommend and aggregate related content, using the relation metadata category as defined by SCORM. We developed a computer system prototype which applies the proposed methodology on a sample of learning objects generating results to evaluate its efficacy. The results demonstrate that the proposed method is feasible and effective to produce the expected results.

Keywords: SCORM, automatic content recommendation, learning objects, information retrieval, text mining.

1. Introduction

SCORM has become the most used international standards related to content for e-Learning and its acceptance is due to the fact that it brings together different standardizations from different institutions, enabling a wide range of applications with reusable, adaptable and easily portable content between the LMS that have implemented its specifications (Su *et al.*, 2006, Rey-López *et al.*, 2009). SCORM defines a Content Aggregation Model (SCORM – CAM) (ADL, 2009a) based on the Learning Object Metadata (LOM), developed by the Institute of Electrical and Electronics Engineers

(IEEE). However, manually dealing with its extensive and complex metadata model is a difficult annotation process (Margaritopolous *et al.*, 2007), which frequently results in insufficient or incorrect metadata (Edvardsen *et al.*, 2009), compromising the quality of learning objects (LO) and restricting the use of the resources provided by SCORM. Among these, the extensibility of its main content by the indication of related LOs is an important example.

The "relation" metadata category classifies different forms of relationships through the following vocabulary contents: ispartof, hasPart, isversionof, hasVersion, isformatof, hasformat, references, isreferencedby, isbasedon, isbasisfor, requires, isrequiredby. In SCORM-CAM there is a definition of how this category is organized, but there is no model established for their use, contrary to what occurs with other categories. It is indicated in the document that the use of metadata, as well as the way it will be done, is defined by content developers themselves or by each LMS (Advanced Distributed Learning, 2009a). Therefore, there are two important issues to note: the inherent difficulty to the manual annotation of metadata and the lack of a specific definition for the use of the relation category in order to build content for e-Learning from related LOs. Thus, this study sought to provide a solution to the question: is it possible to establish a methodology for automatic recommendation of related content, which enables to identify and establish relationships between LOs, by relation category, without having to change its metadata, to modify the standard SCORM or to use custom implementations in LMS?

The following hypothesis was the basis for the development of this work: the establishment of relations between LOs, as defined by the relation category SCORM without changes, is possible to be done from a self-recommendation system that relates contents, using information retrieval and text mining strategies. For this, one should adopt a methodology that employs a domain knowledge base, which is capable of modeling and conceptualizing the field of knowledge relevant to the LOs, so that their contents can be characterized from the main concepts presented in them and that the relation can then be established between the content of one or more LOs.

This paper presents the results of an experimental research to evaluate a methodology that uses ontologies, automatic metadata annotation, information retrieval and text mining to identify and aggregate related content, using the metadata of the category "relation" as defined in their specifications. A prototype of a computer system was developed, which applies the methodology proposed on a learning objects sampling and generates the results necessary for the evaluation of its effectiveness against the problem presented. The results, which analyzed and evaluated with the support of educators, who work in the development of content for e-learning, demonstrate that the proposed method is feasible and effective, producing the expected results.

This paper is organized as follows. Section 2 presents a review of the related work. Section 3 presents the methodology for the recommendation and aggregation of related learning objects proposed in this paper. Section 4 presents the system implemented to evaluate the proposed methodology. Section 5 presents tests and results. Section 6 presents discussions and section 7 presents the conclusions and future work.

2. Related Work

Engelhardt *et al.* (2006) present in their work an approach which is guided by the use of a set of metadata, based on a subset obtained from LOM plus another set proposed by the authors, to establish semantic relations between various LOs present in a repository. These relations were formalized by means of an ontology developed on the basis of the Web Ontology Language (OWL), applicable to a set of inference rules, establishing a semantic network for the entire repository and creating links between the LOs in order to express the relations identified between them. Such relations are established when a new LO is inserted into the repository, so when accessing a LO, the student finds several possibilities to browse through related content by means of the interconnection among the various LOs that make up the network. However, when thinking about processes that require the definition of a unit of learning, course or training, whose content is specific and demands an ordering of concepts and pre-set topics, based on a didactic and pedagogical planning, the solution proposed by Engelhardt *et al.* (2006) may not be adequate. It provides mechanisms that lead to a very dispersed access to a whole variety of LOs in a repository, since all the documents are previously interrelated and the access to the repository is made from a query by the student himself. This can cause a very large number of navigation possibilities, because the relationships are indicated for each accessed LO. As Lu *et al.* (2010) and Edvardsen *et al.* (2009), Engelhardt *et al.* (2006) use an extension of the metadata model defined by LOM and a specific LMS to carry out their work. However, there is no approach that can use metadata to classify and correlate LOs without any interventions in SCORM, and without the need to use a specific LMS to do this, thereby ensuring portability and compatibility of the LOs in different LMSs.

Edvardsen *et al.* (2009) proposed an approach that used the automatic generation of metadata to assist in the organization of didactic and pedagogical contents, aimed for building LOs in accordance to SCORM. In this way, they developed a framework for generating metadata in accordance with the LOM from a given LO, obtaining as a result a new LO, but in SCORM format. For this, they used recovery processes of contextual metadata based on information present in the LMS of the University where they work, in a catalog of courses offered by the university and in entities extracted from the LOs, combining different approaches for the automatic generation of metadata from their respective content. All these metadata was finally referenced by LOM elements, but their sources denote heavy reliance on external elements of the LOs and on how these data are presented, because, mostly, they were not extracted directly from the contents of LOs. Despite believing that their research obtained inconclusive results regarding the quality of obtained metadata, the authors concluded that a LMS can be used not only for the publication of LOs, but also as a source of what they called contextual metadata, that can be used as basis for the generation of specific metadata to the LOs. This approach can restrict its scope of application because it depends on a framework developed to use information that may not be found in other systems and institutions, making difficult the retrieval of the same set of metadata in another context. Therefore, when thinking about content repositories, the large volume of metadata obtained by Edvardsen *et al.* (2009)

can assist in the storage, classification and recovery of LOs. However, there is no clear proposal for use of this metadata in a sense that goes beyond the registration or consultation of this information.

Roy *et al.* (2008) also extend the metadata provided by LOM, in their educational category. In their work, the authors defined a strategy that uses an automatic annotation of LOs, available in content repositories, in order to enable the LMS a proper selection of learning materials, and facilitate the work of content developers in the reuse of this material. In this sense, they developed an ontology whose attributes could characterize the learning materials from a pedagogical point of view. This structure composed its domain knowledge base, which was hierarchically organized into three layers, named respectively as term layer, concept ontology and topic taxonomy. Several terms present in the first layer were associated with sets of concepts that make reference to them, presented in the second layer. These concepts allow the identification of issues related to the topic taxonomy layer. Both the ontology model, developed by the authors, as the strategy for automatic annotation of LOs defined by them, appear as a consistent and workable proposal for selection and reuse of LOs. However, as we can see in Lu *et al.* (2010), Edvardsen *et al.* (2009) and Engelhardt *et al.* (2006), the authors used metadata that extend those who are provided in LOM educational category. Moreover, their approach offers good mechanisms that allow the classification and recovering of LOs in repositories, but it is not intended to establish relationships between them.

According to Nauerz *et al.* (2008), one of the greatest difficulties faced by Internet users consists in finding relevant content to their research, as users need to search for "background information", that is, contents that offer them additional or complementary information for better understanding of what in fact they are searching. Therefore, seeking for a recommendation system based on the user interactions in Web systems, they proposed a framework for annotation of related contents, in files of type eXtensible Hypertext Markup Language (XHTML), which uses data analysis unstructured services, called UIMA and Calais, to automatically analyze a given content and identify certain terms capable to describe certain types of "entities", referring to persons, locations, companies, among others. Thus, the relationships between the contents are described using semantic tags that contain such entities which, in turn, can be linked to related services, as in the case of an entity of type "local" and the Google Maps service. In a previous experiment, the authors argued that their system has produced a large number of irrelevant recommendations, because the processes of generation of semantic tags were performed directly on the content without taking into account user interests and preferences. To solve this problem, they introduced a user model that provides data about their interests and, from this model, selected the fragments of information to generate the tags, allowing the identification of related content whose information was really of interest. This recommendation process, similar to what occurs in Engelhardt *et al.* (2006), may not provide the necessary support when the goal is to build a cohesive unit of learning, whose content is composed of grouped LOs not only because they are related, but in accordance with an organizational structure of pedagogical nature.

It can be seen that the problems raised by Roy *et al.* (2008), Edvardsen *et al.* (2009) and Lu *et al.* (2010), among others refer mainly to the difficulty of identifying LOs

whose contents are in any way related. The possibility to create relations between contents, so that a LO can reference others from a repository, can be thought in a way that references are available directly from each of the LOs that comprise the same content in a SCORM package. Thus, each LO would have, in a specific field, a list of other LOs that could be related to it. This was the premise from which Lu and Hsieh (2009) proposed the use of the relation category from the SCORM - CAM. However, for the authors, the relations described by the relation metadata category are limited, because they can only describe relations guided by the structure of the content, not being able to establish semantic relations between LOs. Given this context, the authors developed an outreach model to the relation metadata category. After concluding their extension model, Lu and Hsieh (2009) obtained fifteen new relations. Then, the usefulness of these new relations, regarding students learning, was tested and analyzed. They stated that the results of their experiments indicated that the new relations were considered useful for most of the 145 students who contributed with their research. Based on these results, the authors considered interesting to draw a common set of metadata and that authoring systems should be created with support to the new model.

In face of the premises indicated by Lu and Hsieh (2009), it's important to highlight a fundamental question about the SCORM and one of its primary objectives: to ensure portability and the reuse of its content packages in any LMS that implement its models. Thus, the development of a set of metadata, which requires a specific implementation in a LMS, is a solution that goes against an essential premise of the standard itself, because it reduces the portability of the content package and their compatibility to other systems. However, as can be seen, many research papers related to SCORM adopted this strategy.

The metadata extension model, prepared by Lu and Hsieh (2009), was applied effectively in the prototype of an LMS developed by Lu et al (2010). As a result, several changes were made in the XML files that do the aggregation of content, inserting different elements and attributes, which were created and interpreted specifically for the metadata model and to the management system developed, respectively, by Lu and Hsieh (2009) and Lu et al (2010). After, the authors proposed a new aggregation format, but without making clear if the LMS' prototype, that they developed, implements support to the rest of the SCORM standard, which consists of SCORM-SN and SCORM Runtime Environment (SCORM – RTE)(ADL, 2009b), beyond the SCORM-CAM itself. Despite having an extensive and rich study about the establishment of relations between LOs and which metadata models are effectively possible to do this, the authors obtained a model far from the SCORM, handled by a system that does not support the standard itself, but rather a specific set of definitions which will not find support in other LMS.

Hernández *et al.* (2009) adopted the following definition of LO: "[...] digital educational material, self-contained and reusable, which has information able to describe its content (metadata)". In this sense, a LO can have fine granularity or coarse granularity. The fine granularity is characterized by smaller content, more objective, such as an example or exercise. The coarse granularity, in turn, is attributed to LOs with more extensive content, that contain other contents, as in the case of an entire course. Based on this

principle, Hernandez *et al.* (2009) have developed a tool they called Looking4LO, which uses natural language processing and ontologies for information retrieval in documents in order to extract LOs with fine granularity from different sources.

According to Hernández *et al.* (2009), the ontology is a critical factor for successful extraction of LOs, because it provides the domain model to an area of knowledge upon which one wants to find and extract content from a document, while the use of a pedagogical model allows one to define what type of content is being sought, that is, exercises, examples or others. Thus, the system developed by them receives as input these two models and a document source. The system output is a set of LOs extracted from the documents, whose content is within the domain model and belonging to one of the elements defined by the pedagogical model, according to the LO metadata. Even getting satisfactory results, the ontology used was very limited, containing only one or two levels of classes and whose instances allowed to carry on correspondences on the documents of the sample used in the tests. According to the authors, the metadata annotations based on ontology did not use information contained in the relations between the entities and did not employ relevant part of the potential offered by domain ontologies. For them, the identification of the semantic relations between the concepts present in the ontology could greatly improve the accuracy of the searches carried out by the system. Furthermore, the identification of these relations could be used to improve the definition of LOs. Their work resulted in a system that can assist content developers in creating LOs with fine granularity. However, the variety of extracted LOs may be large or small depending on the variety, size and quantity of the documents present in the source of content provided as input to the system, which may result in redundancy of these LOs, causing, at the end of the process, a low performance. In addition, verification and selection of LOs generated in the output is up to the system user who wants to use them in the composition of an LO with coarse granularity. Also in this sense, Looking4LO does not provide features that will assist the e-Learning content developer in the composition of a more extensive and complex content, as in the case of a course, that requires the identification and selection of interrelated LOs, that are able to provide complementary and sometimes sequential contents that compose, as a whole, the final content.

Maratea *et al.* (2012) sought to realize the automatic extraction of metadata defined in the general category, contemplated by SCORM - CAM, to classify LOs composed of scientific articles. As some of these metadata, according to the authors, are closely related to the structure and sections of the document, as in the case of the title and the description, and others are evaluated from their own content, such as language and coverage, different techniques were implemented for each metadata type. For metadata extraction considering structural information, a preprocessing step on each PDF file was applied, obtaining, for each one, an XML file that separates and organizes each section of the document. This resulting file was then submitted to an analysis strategy based on rules for extracting the metadata. The Vector Space Model was used as a strategy for natural language processing. For this purpose, all of the documents layout information was removed. From the tests applied on a set of 17 scientific papers, Maratea *et al.* (2012) found that the techniques proposed by them allowed the correct extraction of metadata, with a good accuracy level. Therefore, they proposed, as future

work, the extraction of more complex metadata and in documents less structured than those used by them.

Huynh and Hoang (2010), in turn, sought to relate scientific articles based on metadata extracted from PDF documents available on the web. According to them, based on the obtained metadata, it is possible to know the documents in which a given article is referenced. To do this, they developed a system that uses information about the document layout, rules built from models and an ontology, which they had built, for papers related to computing. The authors point to the fact that, in the proposed approach, care must be taken when creating rules and models, and that the survey of several models consists of a laborious task that requires time and domain knowledge. Thus, they propose as future work to combine their current methodology with the use of machine learning algorithms, in order to increase their accuracy and extract new metadata groups based in Dublin Core Metadata. Also in this sense, the authors state that the creation of rules and models for metadata extraction in bibliographical references, together with the identification of the relationship between them, could help user to identify documents which reference each other, as well as to check if a given reference is valid. After a brief presentation of the main steps performed by their algorithm for metadata extraction, as well as showing some examples of rules defined by them, Huynh and Hoang (2010) did not make it clear in their article how the obtained metadata can be used for organizing documents in digital libraries. The same can be said regarding the use of these metadata for identifying relationships between different articles.

A very similar strategy was used by Guo and Jin (2011b), when they present a system called SemreX. It is a peer-to-peer (P2P) system for sharing text documents between researchers in computer science, that implements a framework based on rules for extraction of metadata related to the title, authors, abstract, periodicals, volume, year and pages present in citations and references from scientific articles. The PDF files, from which metadata is extracted, are converted by the system in two different formats: a simple text file and a XML. The text file contains all the text of the source file, but without layout information. XML, in turn, uses spatial references from the source document to reference the blocks of text and then, for each one of them, store the layout data. For the authors, the formatting information helps to identify the type of content and assists in the extraction of metadata, making the process more accurate. From thereon, the authors apply algorithms based on rules, with the use of knowledge bases, for extracting metadata and subsequent update of the knowledge base used.

The approaches proposed by Huynh and Hoang (2010), Guo and Jin (2011a) and Guo and Jin (2011b) consist of efficient ways to extract metadata from scientific articles. However, they exploit the structural aspects of the documents as a primary reference for the adopted strategy. Such approaches may not be effective when the set of documents is heterogeneous, as in the case of pedagogical contents that does not necessarily have a standardized structure for the presentation of its contents, as occurs in scientific articles.

According to Tuarob *et al.* (2013), DataOne consists of a data network built to facilitate access to data about ecological and environmental sciences worldwide. These data are obtained from different providers and made available through a search interface

called ONEMercury. However, the set of keywords used in searches by users is preset and can be changed only by system administrators in order to avoid the appearance of invalid keywords, because this set is used for manual annotation during the survey process data which, as seen, are derived from different sources. Thus, the problem lies in the fact that, in this way, it is necessary to deal with different annotation levels on data obtained from different sources, of which many may contain meaningless annotations to the ONEMercury, causing data to be lost during searches. Therefore, the authors present algorithms they developed for automatic metadata annotation. Their strategy consists of transforming the problem of annotation in a tag recommendation problem, based on a keyword library, like the one understood by the ONEMercury. In short, the poorly recorded metadata in the files analyzed, with respect to the set used by the DataOne search interface, are again annotated with similar metadata, resulting in a new annotation and decreasing the chances of not being considered on user research. As can be seen, the problem presented is mainly due to the diversity of data sources employed by the DataOne, as they need to be surveyed from a single interface that uses its own library of keywords.

Techniques of IR were employed by Tuarob *et al.* (2013) in their algorithms. However, these require, as pointed out by the authors, a large training over the keywords library, which can be expanded and modified at any time by the system administrators, requiring new training. In addition, the tag recommendation process should be better evaluated, according to them, with regard to its efficiency and scalability.

According to Dorça *et al.* (2016), students tend to have better performance when studying customized contents according to their preferences. Therefore, the authors propose an approach that classifies and filters LOs according to the student Learning Style (LS), by means of an expert system that implements a set of rules and automatically recommends the best adapted and ranked LOs to the student, based on a dynamic studying modeling approach and considering the IEEE LOM. They also emphasize the need to perform automatic correlation between LSs and LOs due to the high difficultly in manually obtaining the best LO to each student. So, the automatic recommendation of LO to LS covers a relevant aspect of content recommendation, but the LOs themselves, principally inserted in large repositories, have their own composition as an important issue. Thus, if on one hand we must deal with the selection of the best LO for a student, as exposed by Dorça *et al.* (2016), it is necessary to treat, on the other hand, the proper composition of this LO, which also consists of a work that is difficult to be carried out by educators, content writers and tutors who work in the elaboration of content for e-Learning. Similarly, ensuring the accessibility, reusability and interoperability of LOs is an important factor in this sense and requires mechanisms capable of describing, classifying and relating them, such as the ontology model and the application profile of IEEE LOM proposed by Solomou *et al.* (2015). Such as Dorça *et al.* (2016), Tarus *et al.* (2017) propose a system to recommend learning resources to learners on a hybrid approach that uses ontology and sequential pattern mining.

To assist in the processes inherent to IR there are several available tools and frameworks. Lipinski *et al.* (2013) presented an evaluation of different approaches and tools for metadata extraction from headers of scientific articles. Nauerz *et al.* (2008) proposed

a framework using UIMA and Calais. Engelhardt *et al.* (2006) used the JENA framework. Maynard (2008) presented a benchmarking of automatic text annotation tools, concluding that the GATE achieved the highest overall rating. Among the surveyed studies, the ones that are more closely related to this research are those grounded in the standard SCORM and dedicated to extract metadata from LOs based on their Content Aggregation Model. Among these, some also sought to establish relations between the LOs based on metadata defined by the relation category, but they ended up proposing extensions to the metadata and a specific LMS to apply these extensions, which reduces the portability and compatibility of the content package in other systems.

3. A Methodology for the Recommendation and Aggregation of Related Learning Objects

The diagram in Fig. 1 shows an overview of our methodology for recommendation and aggregation of related LOs, in accordance with SCORM, as proposed in this paper. It covers three stages, briefly described in the following sections.

The first step consists of relevant information retrieval to each of the LOs. Thus, from a knowledge base represented by (2) in Fig. 1, a set of LOs (1) is subjected to a process of automatic metadata annotation (3), that identifies and classifies its key terms and relevant concepts. Then a hierarchical classification of these terms and concepts is made regarding their level of relevance (4). The LOs, properly annotated (5), are then stored in a repository (6), from which they can be selected (7) and used for the composition of a given content. This content is subjected to a related content recommendation process (8), from the documents in the repository. In this process, other LOs are searched in order to be aggregated as related content and, in the end, the recommended documents can be kept or deleted manually (9). Joining preselected documents and recommended documents, a content package in SCORM format is generated according to the specifications of SCORM-CAM (10).

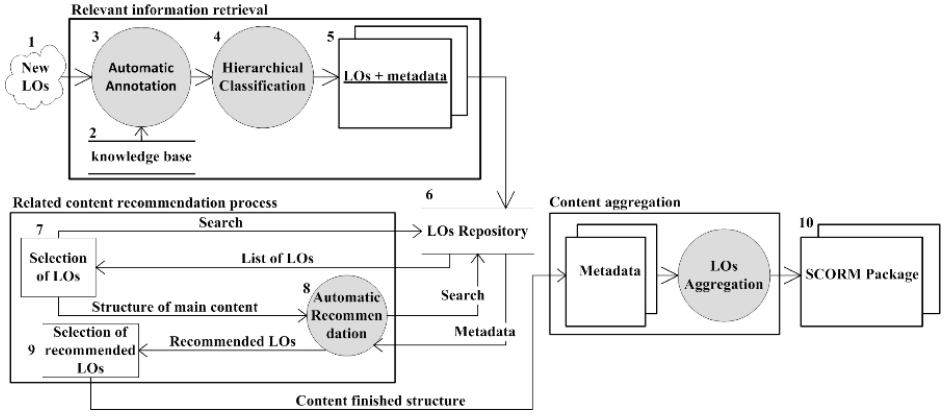

Fig. 1. Flow of the processes to be performed.

3.1. *Relevant Information Retrieval to a LO's Content*

When building a particular course or class, the contents of the used LOs are associated with a particular knowledge field. Therefore, it is necessary to retrieve information, from their content, that can represent it as a whole, summarizing the main issues addressed in it, and that are relevant to the knowledge field to which it is related. Thus, in a text that deals with the biography of an important composer of classical music, for example, it is not interesting to identify any names of people or places that occurs in it, but rather those that are related to the knowledge fields comprised by the classical music domain, so that they can then be further analyzed regarding their relevance to the content of the document itself. These elements will make up the set of the most relevant terms and concepts to the document, which characterizes it as the content present in it.

In this context, it is necessary to have, as a primary reference, a domain model able to characterize and represent the knowledge area to which the LOs belong, over which it intends to apply strategies for information retrieval that, in this case, becomes a relevant information retrieval process. Thus, as discussed in the literature, the use of a domain ontology is fundamental.

Given this, a strategy for information retrieval was defined, which uses a domain knowledge base, made up of a domain ontology and a dictionary of terms that contemplates this ontology. This strategy consists of two main stages. The first stage consists of the generation and automatic annotation of metadata to identify key terms and relevant concepts, from the domain knowledge base. The second consists of the analysis of previously annotated metadata, seeking the hierarchical classification of the elements identified by them, such as their relevance regarding the content as a whole. At the end, the result is a list of terms and concepts, ordered by their degree of relevance. Both stages are carried out during insertion of new LOs in the content repository, which comprises four main processes in accordance with Fig. 2. The first two consist of the loading of new LOs and of the automatic annotation for each of them, comprising the first stage. Next, as part of the second stage, which uses the output generated by the previous stage, a hierarchical classification process of the annotated elements is performed, followed by the storage of the LOs, along with their metadata, in the content repository. Each of these processes will be detailed in the following.

To be loaded, the new LOs need to be inserted in a corpus of documents, which receives a name, so that they can be grouped and identified within the repository, enabling their easy recovery for composition of a given content. Thus, an existing corpus can be loaded from the repository, or a new one can be created.

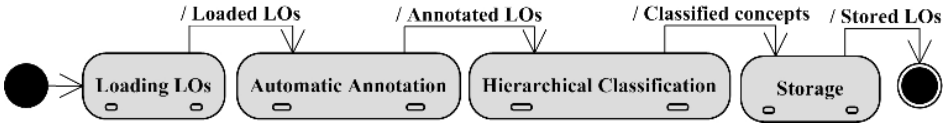

Fig. 2. Loading, annotation and storage of new LOs.

Once loading the new LOs in their suitably corpus, the following process consists of the automatic generation and annotation of metadata, on each of the documents. In order to optimize the identification of symbols, terms and concepts that may be annotated, it is important that only the textual data inherent in the document content and those relevant to their identification, such as authors, title and keywords are present. Thus, the first process to be performed is the removal of the marks that are not standard of text layout and structuring, as well as marks and annotations inserted in the document by authoring and editor softwares, or any others that do not belong to a standard input set. Thus, in the case of LOs in HTML format, for example, the text and the proper text markup language elements would be maintained, and any other data not belonging to this set would be removed, avoiding their processing in the following steps, which could affect the desired output, besides unnecessarily increasing the processing time of each document.

After removing the marks, it is also necessary that the different symbols present in the document be separated and identified as numbers, punctuation symbols or words. This is an important preprocessing stage for other processes because, besides the metadata related to the domain knowledge base, other metadata groups must also be annotated in the following steps, with respect to the grammatical class of a term within a sentence, or even for each sentence recognition, for example. These processes use up the application of different sets of rules, lexicons and data from the knowledge base to process each symbol on the document, which shows how much a correct separation and identification of these affects the efficiency and accuracy of the process of generation and annotation of metadata as a whole.

Once the several symbols present in the document have been identified, the automatic process of generating and annotating metadata for identifying the terms and concepts relevant to the content of each LO begins. All terms and concepts in the document that are defined in domain knowledge base, must be properly annotated with its kind, taxonomic position and classification according to the ontology. In addition, it is also necessary that they receive annotations related to its grammatical class, considering that nouns and proper nouns, for example, may have greater potential of relevance than others, such as adjectives and adverbs. Thus, from the beginning to the end of the document, each line needs to be examined, applying a set of rules in order to identify and separate each of the sentences, so that the grammatical class of a term or concept may be identified by an analysis of its syntactic position within the sentence to which it belongs. So after the text subdivision into well-defined sentences, their terms and concepts can receive annotations with metadata identifying its grammatical class. Those who are not properly identified are annotated as "unknown". All these metadata are extremely important so that it will be possible to submit one or more LOs to the process of automatic recommendation of content.

At the end of the process, it is necessary to seek references for symbols that have not been identified, but may be related to important terms and concepts for the content of LO. For that, matching algorithms should be applied, so that each term annotated as "unknown" can receive the same annotation given to a corresponding term that was previously annotated.

Once a corpus, present in the content repository, has passed through the automatic metadata annotation process, the hierarchical classification phase of their key terms and relevant concepts starts, in which it is necessary to determine which of these elements are the most representative regarding the pedagogical content of each LO. For this, each one of them needs to be analyzed in order to assign them a value that allows measuring its relevance to the content of the document and in accordance with the domain knowledge base settings.

The techniques used at this stage, to the calculation of the relevance of each term annotated, were defined based on the literature on information retrieval and text mining. As described in the literature, for information retrieval there are some techniques that consider the relations of the terms and concepts of a given text with all text found in a collection of documents. These were not used, because they are applied primarily to the calculation of similarity between documents, or between them and the input data from a query, provided by a user (Morais and Ambrósio, 2007), whereas the aim of this study is to establish relations between different documents, seeking not the similarity, but the completeness between their contents. Therefore, techniques were employed that allow analysis of the terms and concepts in relation to the document itself.

The goal is to associate, to previously annotated metadata, metrics about the frequency of the term in the document and about their potential relevance, based on their position in the document structure, in the sentence where it occurs and in its grammatical classification. To each metric are calculated the final relevance of the term under review, which goes also to be associated with it as a new metadata. The relevance final values are used to generate a vector of relevance, which contains the terms better weighted, ordered by the greater weight, representing the hierarchically most relevant information to the document contents to where they belong. This vector of relevance will be given as input to the process of automatic recommendation of related contents.

After the processes of loading, automatic annotation and hierarchical classification of relevant terms and concepts, the LOs contain all the information necessary for the subsequent composition of a given unit of learning and submission to the related content recommendation processes. Thus, the corpus where they are located can be closed and stored in the content repository.

3.2. *Automatic Recommendation and Aggregation of Related LO*

Since there is a repository of LOs, properly indexed and that have been submitted to the processes for relevant information retrieval as defined in the previous section, these LOs are ready to be submitted to the searching and related content recommendation processes, according to the strategy defined in this research. Its main processes are shown in Fig. 3.

Fig. 3. Recommendation and aggregation of related LOs – main stages.

The first process consists of selecting LOs to compose the main content of a SCORM package. Then the automatic recommendation and aggregation of related content process is performed, in conformity with the relation category. Finally, the user responsible for the development of pedagogical content can then select from among the recommended LOs, those who actually will be aggregated to the main content and inserted into the SCORM package. So, it is an automatic recommendation process, followed by a semi-automatic process to aggregate the selected content.

Once the user has accessed the repository of LOs and selected those that will compose the main content, the list of them is given as input to the next process, which will perform the automatic recommendation of content related to them, from the content repository. The structure of the main content is covered and each of their LOs have their metadata analyzed in search of their relevant concepts, prior and duly annotated and classified hierarchically into the vector of relevance. For each identified concept a relation with their classification in the domain ontology is established, which, in turn, describes a graph in which the classes of concepts define its vertices and the relations between them are defined by their edges. From these relations, between an annotated concept and its ontological class, other relations that can be established between one concept and the others present in other classes of ontology are identified. Thus, for each identified ontological relation, an association in the format **(LO, relevant concept, relations)** is generated, where the number of associations for a LO is defined by the sum of the associations generated for each of its relevant concepts, when each of them can have more than one relation depending on the number of edges in the vertex that defines its ontological class. The generated associations are displayed to the user, while the process is repeated for the next relevant concept, until all the vector of relevance is covered, or until the user decides to stop the process of analysis and jump to the next LO.

Once all LOs that composes the main content has been analyzed, the various associations of concepts that were generated for each of them are then transformed into recommendations of related LOs. The recommendations are listed so the user can select those he wants to keep. Then, they are stored so that they can be used in the packaging process, which is responsible for aggregating all of the contents, in accordance with the SCORM-CAM, and generating the content package in SCORM format. The process of generating recommendations, goes through the list of LOs of the main content and retrieves, for each of them, the associations previously generated and stored. These are used to generate a set of recommendations for each of the LOs, from the concepts and relationships contained therein. For each association of a LO we recovered: its relevant concept, the relations contained therein and, from these, the ontological classes pointed as associated with the relevant concept. From these ontological classes, a search is performed in the content repository for LOs that belong to them and have, among its most important concepts, the concept in the association being analyzed. For each LO found in the repository that matches these criteria, a recommendation of this LO as being related to that LO of the main content is then generated, for that type of relation, and has this format: **Recommendation(LO, relation(concept, recommended LO)).** All recommendations generated are listed, allowing the check of the title, abstract, keywords and

authors of each LO. At this point, the user can select the recommendations he wants to keep to the composition of the final content of the SCORM package.

The packaging step includes the process responsible for the aggregation both of the main content as the recommended content, generating the SCORM package ready for publication in an LMS. At this stage, one list of LOs is generated from the main content and given as input to the packaging process, where the *imsmanifest.xml* file, that is responsible to describe the SCORM package content, is written to store the access information to the content through the SGA, as the specifications of SCORM-CAM. For each LO in the list, the metadata related to its title and the URI containing their addresses are retrieved from the repository, so that the reference in *imsmanifest.xml* can be created. After the aggregation of all the main content, the list of LOs is then processed to the aggregation of the related content. For each LO from the main content the recommendations previously generated for each of its relevant concepts are then retrieved. These recommendations are analyzed and the recommended LOs also have their title and URI retrieved from the repository, becoming to be associated with the main LO, so that the LO on the recommendation is associated by means of the appropriate attribute of the relation category and its reference is, then, inserted in the *imsmanifest.xml*. At the end, all the LOs are copied from the content repository and encapsulated together with the *imsmanifest.xml* file, composing, finally, the SCORM content package resulting from the entire process.

4. Related-Content Aggregation and Recommendation System

In this section, we present our strategies and approaches to address the methodology steps presented in Section 3. We used the IDE NetBeans to prototype a related-content aggregation and recommendation system. In particular, we used GATE plugins, presented in Cunningham *et al.* (2012), to develop a Java application with four modules: i) AssignerRelevance, for automatically metadata annotation and hierarchical classification of relevant concepts; ii) AssociationsBuilder, to generate relations between the relevant concepts and the domain ontology; iii) RecommendationsBuilder, to generate related-content recommendations from the pre-established relations, and; iv) DocScoreRecommendationsBuilder, to score and rank recommended documents.

Section 4.1 presents the domain knowledge base with the dictionary of terms and the domain ontology. The sections 4.2, 4.3 and 4.4 discuss the processing steps and present the implemented modules.

4.1. *Domain Knowledge Base*

In this work, we use the classical music as our domain knowledge area. Particularly, we built a knowledge base with a dictionary of terms and domain ontology for classical music, using them to automatically annotate terms and concepts. Terms in dictionary are grouped in ontology classes interrelated by links and taxonomic relationships. The

dictionary contains 37.183 terms and concepts, distributed in 47 ontology classes. Fig. 4 shows an excerpt of the taxonomic classification file "lists.def".

The OntoMusica ontology was a candidate ontology for the classical music domain area. However, such ontology contains few classes of concepts with a few relations between them. Thus, we proposed an ontology with 39 classes and 31 relations between them. To model our ontology we used the Unified Modeling Language (UML) and to build it we used the GATE's ontology editor, as shown in Fig. 5 and Fig. 6.

```
musical_genre.lst:musical_genre
musical_genre_chanber.lst:musical_genre:genre_chamber
musical_genre_opera.lst:musical_genre:genre_opera
musical_genre_orchestral.lst:musical_genre:genre_orchestral
musical_genre_solo_instrumental.lst:musical_genre:genre_solo_instrumental
musical_genre_vocal.lst:musical_genre:genre_vocal
musical_genre_orchestral_ballet.lst:musical_genre_orchestral:ballet
```

Fig. 4**.** Part of the file "lists.def".

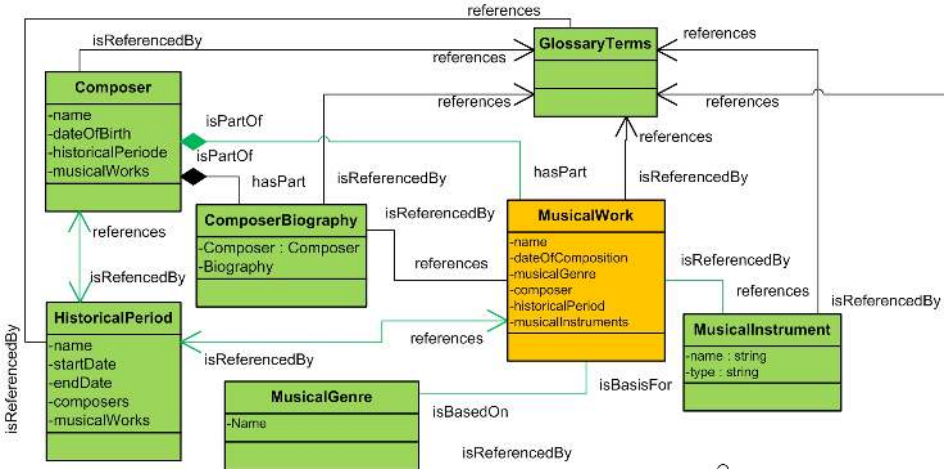

Fig. 5. Modeling domain ontology.

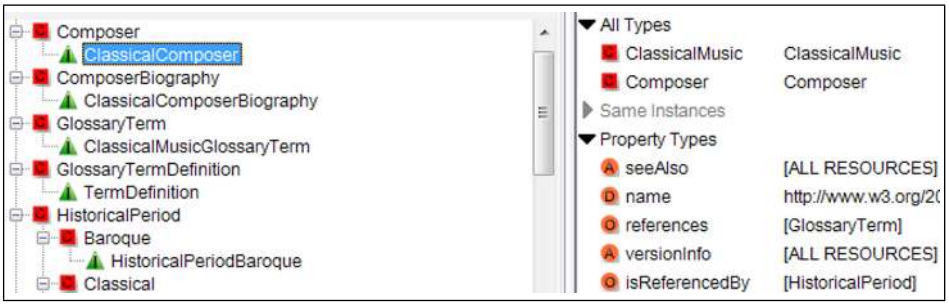

Fig. 6. Building ontology.

Different from Lu and Hsieh (2009), in our ontology the relations between classes are not based on the document structure, but are based on the topic organization presented by NAXOS and in the text characterization from their central theme, which is suitable for documents that consists of LOs, i.e., a unit of learning whose content presents a certain subject and ends in itself, that may or may not be extended, but in any case, being able to be understood by itself (Hernandez, 2009).

Thus, using a class diagram, it was possible to map the relations defined in the relation category for the relations established in the UML, through associations, aggregations, inheritances and specializations. This approach allowed confer a semantic character to the metadata from relation category, when having, for example, two classes denominated Composer and ComposerBiography, whose association takes place through an aggregation, with which it is established that Composer has a relation of the type *haspart* with ComposerBiography and this, in the opposite direction, establishes a relation of the type *ispartof* with Composer. Considering that the class diagram was used for modeling a system for cataloging composers, the aggregation relationship between a composer and his biography would be coherent. Likewise, it was observed that the *haspart* and *ispartof* relations may contain the same meanings denoted to the aggregation in the UML. In another example, a musical work is part of a composer, as an aggregation, in which if a composer ceases to exist, his works too. Therefore, there is an aggregation of the type *haspart* and *ispartof* between the classes Composer and MusicalWork. The same was seen for the other relations and their respective associations in the UML. Between composer and historical period, it can be considered that a composer is associated with a period, but if this period ceases to exist in the study of history, the same does not occur with the concrete elements that were associated with it. Thus, if a composer was associated with an historical period, and this ceased to exist, the he can be associated with other period. Thus, it was adopted that a composer has as a reference a historical period and a historical period make reference to a composer, receiving the *references* and *isreferencedby* associations. In turn, the understanding of music history requires an understanding of each of its periods, to the extent that each period can be viewed as a specialization of HistoricalPeriod class. This also applies to musical genres, because the knowledge about all musical genres requires the knowledge of each genre in particular, to the extent that each genre can be viewed as a specialization of the MusicalGenre class. In these cases, there is a relationship of the type *isrequiredby* and *requires*, in sense that the general class requires specialized classes and these are required by it. In the case of musical work and musical genre, if a musical work is characterized as an opera, for example, it means that their composition is based on this musical genre and that this musical genre, in turn, served as a basis for their composition, making an association between the musical work and the musical genre. Thus, there is a relation of *isbasisfor* and *isbasedon*.

The relations receive the following weights for association of the metadata in the SCORM: *requires*/*isrequiredby* and *isbasisfor*/*isbasedon* are strong relationships, in the didactic and pedagogical sense, where a content needs the other to be understood; therefore, the LOs will be interrelated as prerequisites. In turn, *references*/*isreferencedby* and *ispartof*/*haspart* presuppose that there is no mandatory complementary relation-

ship, where documents complement one to another but do not depend on each other to be understood, so the LOs will not be related as prerequisites one to another, but only as supplementary materials. Finally, a map was created through the editor of the GATE, which associates each list of dictionary terms to a class from the ontology. This map is stored in a text file that, along with the dictionary of terms and the RDF file containing the ontology, makes up the domain knowledge base defined and used in this research.

4.2. *Retrieval of Relevant Information*

As discussed in Section 4.1, to be able to recommend and aggregate related LOs, it is necessary that they have undergone a preprocessing step, which aims to retrieve information relevant to their content, enabling the analysis and identification of possible relations between them. To implement this step we used a plugin from GATE framework called ANNIE, discussed in Cunningham et al (2012), and the AssignerRelevance module, that was implemented as part of this work. The ANNIE subdivides the text into symbols and sentences, annotating the terms with their grammatical class and indicating its ontological class. The AssignerRelevance in turn, implements the algorithms to the generation of the other metadata, defined for this step, to which have been associated metrics necessary for information retrieval, as well as performs the hierarchical classification of these elements, from the relevance attributed to each of them.

From the domain knowledge base, the generating processes and automatic annotation of metadata, as well as the hierarchical classification of relevant terms and concepts, are realized upon the insertion of new LOs in the content repository as defined by the proposed methodology. For manipulating the repository, a Serial Data Store was implemented, using the API of the GATE, which serializes and stores the LOs in their respective corpus.

Since new LO, that will be inserted in the content repository, have been loaded by the system from its original repository, the next step consists in the generating process and automatic metadata annotation in each of the documents. For this purpose, the system uses the ANNIE plugin. This takes as input the corpus containing the LOs, the dictionary of terms, their mapping to the ontology and the domain ontology. The output of ANNIE consists of LOs containing terms and concepts annotated regarding its grammatical class and their ontological classification. As a preparation stage of the LOs for the recommendation of related content to them, after the annotation of metadata, it is necessary that the terms and concepts annotated are classified according to their level of relevance regarding the content as a whole. Finally, the LOs are persisted, with the appropriate annotations, in the content repository.

For the hierarchical classification of the key terms and relevant concepts, the system uses the AssignerRelevance module, which was implemented to receive as input a list of LOs and provide as output the same list, but with new metadata inserted into each LO, for each term or concept previously annotated, and a hierarchical classification of their set based on the level of relevance of each them in relation to the text as a whole. The new metadata inserted by AssignerRelevance contain metrics of relevance to each of

Table 1
Relevance indicators and possible approaches

Indicators	Possible approaches
More used words in the text, without stop-words	Calculus of the relative and absolute frequencies of the word in their respective document
Words present in titles, keywords and abstracts	Word position in the different text sessions, from an analysis of the document structure
Words that are substantives and complements	Semantic analysis and identification of the syntactic position of the word
Words that can be defined by others in the sentence	Semantic and syntactic analysis to verify the relations between two words in the sentence. Example: Mozart is a co+mposer

the terms and concepts. From the analysis of the new generated metadata, it is possible to associate, to the term or concept, a certain weight, which can be inferred based on a set of different indicators of relevance. The Table 1 lists some of these indicators with the approaches commonly used to measure them (Morais and Ambrósio, 2007). Using a combination of these approaches, the calculation of the relevance for each annotated term is carried from the formulation proposed in this work, which is illustrated in Fig. 7 and set forth below.

From the formulations shown in Morais and Ambrósio (2007), Roy *et al.* (2008), Tuarob *et al.* (2013), the following formulation was proposed for this work: consider that **VT** is a vector of relevant terms , $\mathbf{R_{ti}}$ is the relevance of a term $\mathbf{t_i}$, $\mathbf{F_{abs(ti)}}$ the absolute frequency of a term $\mathbf{t_i}$, $\mathbf{F_{rel(ti)}}$ the relative frequency of a term $\mathbf{t_i}$, $\mathbf{T_{tit}}$ is a term in the title of a document $\mathbf{d_i}$, $\mathbf{T_{KW}}$ a term present in the keyword set of a document $\mathbf{d_i}$, $\mathbf{T_S}$ is a term that is substantive, $\mathbf{T_{Frel}}$ is the term with the higher frequency relative and $\mathbf{Sent_{ti}}$ the sentence where the term occurs. So, the following functions are defined: (1) returns the absolute frequency of the term $\mathbf{t_i}$ in the document $\mathbf{d_i}$; (2) returns the relative frequency of the term $\mathbf{t_i}$ in the document $\mathbf{d_i}$, where **N** is the total number of terms in the document; (3) receives a term $\mathbf{t_i}$ and the sentence where it occurs and returns 1.5, if it is followed by a noun preceded by a linking verb, increasing its relevance, or 1 otherwise; (4) returns a factor of relevance of a term $\mathbf{t_i}$ for the document $\mathbf{d_i}$, where it occurs, from a valuation based on a combination of parameters in $\mathbf{F_{rel(ti)}}$, $\mathbf{T_{tit}}$, $\mathbf{T_{KW}}$ and $\mathbf{T_S}$; (5) returns the final relevance of a term $\mathbf{t_i}$ for the document $\mathbf{d_i}$ where occurs, using the values returned by (3) and (4) to confirm the weight of (1).

$$\textbf{(1)}\ F_{abs(t_i)}=\text{FreqAbs}(t_i,d_i)\quad \textbf{(2)}\ F_{rel(t_i)}=\text{FreqRel}(t_i,d_i)=\text{FreqAbs}(t_i,d_i)/N\quad \textbf{(3)}\ \text{FuncDef}(t_i,\text{Sent}_{t_i})$$

$$\textbf{(4)}\ R(t_i,d_i)=\begin{cases}2.0, (t_i=T_{Frel}) \wedge ((t_i=T_{tit}) \vee (t_i=T_{KW})) \wedge (t_i=T_s)\\ 1.5, ((t_i=T_{tit}) \vee (t_i=T_{KW})) \wedge (t_i=T_s)\\ 1.0, ((t_i=T_{tit}) \vee (t_i=T_{KW}))\\ 0.75, (t_i=T_{Frel}) \wedge (t_i=T_s)\\ 0.25, (t_i=T_s)\end{cases}$$

$$\textbf{(5)}\ \text{FuncRel}(t_i,d_i)=\begin{cases}F_{abs(t_i)} \text{ x } R(t_i,d_i) \text{ x } \text{FuncDef}(t_i,\text{Sent}_{t_i}), R(t_i,d_i)>0\\ 0\end{cases}$$

Fig. 7. Defined functions to the relevance calculation.

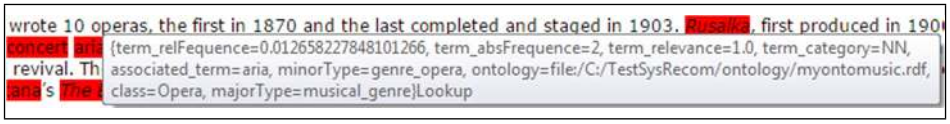

Fig. 8. Annotated word in a LO and respective metadata.

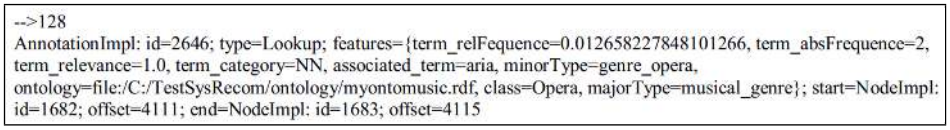

Fig. 9. Part of the log file of the annotations and generated metadata.

Employing the functions defined in Fig. 7, for each term $\mathbf{t_i}$ in $\mathbf{d_i}$, it is calculated $\mathbf{F_{abs(ti)}}$ = **FreqAbs** (t_i, d_i) and $\mathbf{F_{rel(ti)}}$ = **FreqRel** (**ti**, **di**). For every term $\mathbf{t_i}$ in $\mathbf{d_i}$ is set its relevance $\mathbf{R_{ti}}$: $\mathbf{R_{ti}}$ = [$\mathbf{t_i}$, **FuncRel** ($\mathbf{t_i}$, $\mathbf{d_i}$)]. If $\mathbf{R_{ti}}$ >=0.25, **VT** ← [ti, R_{ti}]. The variation between 0.25 and 2.0 for the return of (4) splits, maintains or doubles the initial weight assigned to term by (1), generating its final relevance value.

Fig. 8 illustrates part of an LO, in HTML format, containing the annotated terms and corresponding metadata generated up to this step, visible when the mouse cursor is over one of this terms. Fig. 9 shows part of the generated file for registry of the data generated for each LO.

4.3. *Associations Building*

Once all the previous steps are completed, for each LO present in the repository, the result is a VT associated with each document. The next step consists in identifying the possible relations for each LO, based on the VT elements. For this, the system uses the AssociationsBuilder module, which receives as input the list of LOs present in corpus and provides as output the associations that are possible for these documents, based on the relations described by the domain ontology, from classes to which the terms and relevant concepts in VT are associated. These associations are entered in the form of metadata in each of the LOs. To this, a parser for the domain ontology was also developed, using resources of the GATE's API, that is used by the AssociationsBuilder and was called OntologyParser.

Thus, for each element in VT, its metadata are analyzed and the class to which they are associated in the ontology, previously annotated, is retrieved. From this class, the parser returns the superclass and subclasses associated with it, as well as the possible relations that it establishes with the other classes, which were set during the modeling of the ontology and that obey the vocabulary of the relation category of the SCORM: *requires* and *isrequiredby*, *ispartof* and *haspart*, *references* and *isreferencedby*, *isbasedon* and *isbasisfor*. Each association contains the relevant term, its class, subclasses and relations, which have their class as domain and the associated class as range, forming a graph on the ontology where the classes are the nodes and the relations are the edges that connect them. Once the associations are completed, the LOs contain all information neces-

```
Building terms ontological association to document: ComposerBiography_Antonin_Dvorak.html_00034
Processing 68 annotated relevant terms.
-->
Relevant term: Dvorak
Ontological class: Composer
Ontological super class: ClassicalMusic
Ontological sub classes: don't have.
Searching relations in the ontology:
Relation name: references
Domain class: Composer
Range class: GlossaryTerm

Relation name: isReferencedBy
Domain class: Composer
Range class: HistoricalPeriod

Relation name: hasPart
Domain class: Composer
Range class: MusicalWork

Relation name: hasPart
Domain class: Composer
Range class: ComposerBiography
```

Fig. 10. Part of the log file of the generated associations.

sary for the process of automatic recommendation of related contents. Fig. 10 illustrates part of the generated file to verify the associations annotated in their respective LOs.

After generating the appropriate associations for each LO, the relevant information retrieval step concludes. Thus, the LOs can then be finally persisted in the content repository, along with their metadata, remaining available to content authors who want to use them to compose a unit of learning.

4.4. *Recommendation and Aggregation of the Related Los*

Once a set of LOs was selected from the repository to the composition of a given didactic and pedagogical content, these can be submitted to the process of automatic recommendation of related contents. For this, we used the system module called RecommendationsBuilder, which was implemented in order to receive as input a list of LOs and provide as output another list, containing a set of LOs recommended as content related to the LOs from the input list. So, this process consists to generate a set of recommendations for each LO from the input list, so that each recommendation points to another LO in the repository and identifies the type of relationship established with the LO to which has being recommended.

Iterating over the input list, each of the listed LOs is retrieved from the repository. Among the metadata annotated at each LO, in the previous steps, are the various associations generated from their most relevant terms and concepts, based on the domain ontology structure. Thus, for each association found, the relations that comprise it are analyzed and the classes of terms that they point to are identified. So we have, through

these relations, arches that connect the document to several other classes of concepts for each of its more relevant terms. Thus, for each relation presented in each of the associations generated for each of the most relevant terms, in each LO of the input list, a search is performed in the content repository for others LOs, whose most relevant terms belong to the reach class of the term through the relationship under review. For each found LO its VT is analyzed. If it contains the source term of the association under analysis, this LO is then recommended as related content to the main LO and the type of relationship is described as being of the type described in the association of source term. Fig. 11 illustrates part of the file that records the recommendations generated for the respective LOs.

In order to refine the recommendations generated and prevent a large number of irrelevant recommendations, even if these have been built through all the processes described above, a final processing step is performed to return to the user the list of recommended LOs. It is a process of ranking the recommended documents for each of the associations.

The recommendations previously generated, sometimes, brings more than a recommended document for the same type of relationship, from the relevant term contained in a same association. Therefore, it is important to determine which of these documents are the most recommended. This process is performed by the system module called DocScoreRecommendationsBuilder. This module takes as input the recommendations for each LO under review, generated by the RecommendationsBuilder. From these recommendations, to the documents identified as related to each term of a given recommendation, are given a score that indicates, among these documents, to which of them the source term of the established relationship is more relevant.

It is important to note that, at this stage, the terms present in the recommendations have already been weighted according to their relevance to each document, by AssignerRelevance. So, the process performed by DocScoreRecommendationsBuilder is not to give a new relevance to the concept, but to group different documents identified as related to the LO under review from that concept, that appearing as relevant to them all, and tell to which of them this concept is more relevant. For this, this system module uses a plugin of the GATE called SearchPR, that receives a term and a collection of documents, returning, for each document, a score that indicates how that document is important, considering the term given as input.

Thus, the list of recommendations returned to the user is able to point, for each relationship established, the recommended LO that is more closely related to the LO for which the recommendations were generated.

```
Building recommendations to document: HistoricalPeriod_SUMMARY_OF_WESTERN_CLASSICAL_MUSIC_HISTORY.htm_00055 in corpus
htmlLOCorpus
Processing 363 ontological associations
-->
Relevant term:Ludwig
Relevante term class: Composer

Searching for hasPart relation in range class = ComposerBiography
Relation found:
HistoricalPeriod_SUMMARY_OF_WESTERN_CLASSICAL_MUSIC_HISTORY.htm_00055 term Ludwig hasPart
ComposerBiography_Ludwig_Minkus.html_0003C
```

Fig. 11. Part of the log file to the generated recommendations.

5. Tests and Results

As part of this work it was organized a LOs repository, consisting of 8.967 documents, whose contents are comprised within the field of classical music. For the tests and evaluation of results, the performing of a manual metadata annotation step and indication of relevant terms was necessary, as will be described below. Thus, because of the difficulty inherent in manual execution of these processes, a sample was generated from the total of LOs present in the repository, keeping the same ratio on the percentage of documents for each category. The documents were selected automatically and randomly, as shown in Table 2. Of the 111 documents obtained, 10 have only the name of a composer and a discography with his works. These were not analyzed, resulting in the end in 101 documents.

Four education professionals accepted the invitation to contribute to this research, voluntarily and according to the procedures described in the informed consent form, duly submitted to the Ethics Committee of the University. Each collaborator has been requested, after having properly agreed, to conduct manual annotation of all documents 101 of the final sample, also indicating the terms they thought most relevant among the annotated terms, to each document, from which others documents that were related to them should be recommended, in order to extend or supplement its content. Only three collaborators have concluded the activity within the period provided for thirty days, resulting in 303 annotated documents. Two results with a greater number of annotations were used to analysis the, whose collaborators are identified as A and B. Each document was analyzed and had the annotations manually computed. The terms indicated as most relevant, by each collaborator, were by them listed in a spreadsheet, for each document.

Table 3 presents the manual annotation results for collaborators A and B, including the number of manually annotated terms, the number of relevant terms for the domain, and the overall accuracy achieved by the manual annotation. Additionally, Table 3 presents the percentage of accuracy for each collaborator, and the average of annotations by document classes, allowing a refined analysis of the behavior of the collaborators regarding the annotation process, as will be discussed below. Other important information, which is contained in Table 3, is the final count of terms indicated as most relevant on the sample, either by each of the collaborators or by the sum of the results. The importance

Table 2
Composition of the initial sampling

Classification	Number of documents	Sampling Percentage
Composers Biographies	37	33,3%
Historical Periods	2	1,8%
Glossary Words	14	12,6%
Musical Works	58	52,3%
Total of documents:	**111**	**100,0%**

Table 3
Manual annotation results

Annotations Over the Sample			Coll. A		Coll. B		Accuracy	
Total	Relevant	Accuracy	Total	Rel.	Total	Rel.	**Coll. A**	**Coll. B**
1892	**1231**	**65,06**	**749**	**492**	**1143**	**739**	**65,69 %**	**64,65 %**
Average of annotation by document by class	Composers Biographies		8,43	4,50	20,07	15,50	53,39 %	77,22 %
	Historical Periods		156,0	117,0	123,0	117,0	75,00 %	95,12 %
	Glossary Words		7,14	5,93	5,50	5,43	83,00%	98,70 %
	Musical Works		4,36	2,81	6,46	1,90	64,59 %	29,40 %
Indications of more relevant terms.			**287 terms**		**191 terms**		**478 total**	

of this information lies in the fact that the basis for relations between the documents, according to the proposed methodology, is the vector of relevant terms generated for each document.

The annotated terms and concepts that did not concern the field of classical music were not considered relevant, having as main reference the domain ontology, and those that referred to general information such as "winner of four awards", and "born in Paris". These annotations were applicants in the case of Collaborator A. These terms were also not computed in the counting of terms indicated as most relevant for each document.

Table 4 shows the number of matching results, consisting of the intersection set of the annotated terms by both collaborators, for each class of documents, as well as the terms which have been indicated, also by both, as the more relevant regarding the sample. These data allow an analysis of possible variations on the pattern of annotation adopted by different actors, i.e., the collaborators A and B.

During the analysis of the material, we can see the difficulty in maintaining consistency in the manual annotation process. To annotate documents belonging to the same class, with the same format and standard for availability of information, sometimes a particular set of terms has been marked as relevant, sometimes not, for the same collaborator. For example, in texts relating to musical works, the collaborator A kept the pattern of annotation, always marking the author, the genre of their musical work, date and place of presentation, with little variation, not annotating elements in the text of the synopsis, which perhaps may be the result of this tiring and repetitive activity of meta-

Table 4
Count of matching results

Document classes	Coincident annotations	More relevant terms
Composers Biographies	53	53 coincidences to the terms indicated as most relevant.
Historical Periods	41	
Glossary Words	41	
Musical Works	13	
Total	**148**	

Table 5
Automatic annotation results

Annotations Over the Sample		Total	True Positives	**Accuracy**
Annotated Terms:		**6228**	**4988**	**80,09 %**
Average of annotation by document by class	Composers Biographies	84,93	71,14	83,77 %
	Historical Periods	1213,00	1083,00	89,28 %
	Glossary Words	23,71	23,07	97,29 %
	Musical Works	39,07	26,95	69,98 %

data manual annotation. The collaborator B, in turn, sometimes has inserted annotations in the synopsis, sometimes in the technical sheet, sometimes in the names of the characters of the operas, making difficult to identify the criteria adopted. Furthermore, the amount of annotations of the collaborator B significantly decreased between the first and last annotated documents. In larger documents, she indicated as the most relevant terms only those present on the first page.

After the completion of the work by the collaborators, as well as the analysis of the material produced by them, the same sample was submitted to the Recommendation System and Related Content Aggregation, running up all processes, from the annotation to the recommendation. The results of the automatic annotation are presented in Table 5.

It's possible to notice a difference in analysis parameters, between automatic and manual results, where in the automatic there is no occurrence of non-relevant terms, because of the consistency of the knowledge base and the fact that only terms present in it are annotated. However, attention in this case turns to the generation of false positives. These occurrences are due to difficult problems of the information retrieval and that are beyond the scope of this work, such as the treatment of homonyms and duplication. However, this questions and other interesting analysis will be discussed in the Section 6.

As described in Section 3.1, the association phase does not process all annotated terms, but only the most relevant in each document. About the sample used were generated 3508 ontological associations over 101 LOs. To specifically test the related content recommendation process, was given as input the LO of the sample containing the highest content, it belongs to the Historical Periods class and totals 363 associations. The relationships established by these associations were analyzed by the system and generated a set of 12 final recommendations for this document over the 101 LOs contained in the sample. These recommendations are divided as follows: one LO as isrequiredby; seven as isreferencedby and four as hasPart. The 12 recommendations automatically generated were correct and are presented in Table 6. Fig. 12 illustrates one of the recommendations.

To perform the test, was used a machine with Intel Core i3 processor, with 2.27 GHz per core, 4 GB of RAM and Microsoft Windows 7. The process of loading, storage, automatic annotation and generating associations over the ontology, for all sampling, were run at 1 minute, 43 seconds and 2 tenths of a second. The process of generating recommendations was performed in 5 tenths of a second.

Table 6
Automatic recommendation results

HistoricalPeriod_SUMMARY_OF_WESTERN_CLASSICAL_MUSIC_HISTORY.htm	
isrequiredby	Music_Theor_Online_Music_of_the_20th_Century.htm
isreferencedby	ComposerBiography_Victor_Herbert.html, ComposerBiography_Colin_Matthews.html, ComposerBiography_Antonin_Dvorak.html, ComposerBiography_Henry_Purcell.html, ComposerBiography_Erroll_Garner.html, ComposerBiography_William_Byrd.htm, ComposerBiography_Ludwig_Minkus.html
haspart	GlossaryTermDefinition_Recorder.htm, Charles_Wakefield_Cadman.htm, Gustav_Mahler.html, Charles_Wakefield_Cadman.htm

Relevant term: classical music
Relevant term class: Classical
Relevant term superclass = HistoricalPeriod
HistoricalPeriod_SUMMARY_OF_WESTERN_CLASSICAL_MUSIC_HISTORY.htm_00055 term classical music
isRequiredBy HistoricalPeriod_Music_Theor_Online_Music_of_the_20th_Century.htm_00054

Fig. 12. Part of the generated file to the final recommendations.

6. Discussion

Based on the analysis of the tests performed, it is observed that the methodology proposed in this research is feasible and produces the expected results, with good precision and efficiency, better than those obtained manually. It can be applied to different areas of knowledge to perform the composition of didactic-pedagogical contents, being necessary only the use of a domain knowledge base related to the desired area. However, there are several factors that could be perceived during the tests and the analysis of the results obtained. These factors point to interesting discussions in the sense of seeking strategies to improve the accuracy achieved.

Most of the false positives obtained in the automatic annotation step are due to homonyms of terms present in the contents of learning objects and in the domain knowledge base. Thus, terms were annotated as relevant to the domain when, in fact, they were homonyms of those who would actually belong to the domain, having as reference the knowledge base and the contents of the learning objects in which they are inserted. Therefore, it is necessary to look for possible approaches for the treatment of homonyms in the context of this work. Moreover, according to the literature, the disambiguation of homonyms is a difficult problem in information retrieval, needing relevant efforts to find new approaches to the problem.

The higher incidence of homonymous is due to the composers names and terms widely spread, easily present in contexts outside of domain area, as in cases of “time” and “scale”, that are present in the glossary as “musical time” and “musical scale”, but were responsible for some of false positives. In one of the documents, the phrase “large-

scale composition" had "scale" annotated as a relevant term to the domain and belonging to the Musical Glossary.

In the specific case of the composers, the organization of the names in three lists (one list of full names, one for first names and another for middle names) caused a problem. At times, names such as Michael Tilson Thomas had more than one annotation, with "Michael" annotated as the second name of a composer and also as a first name, "Tilson Thomas" as a middle name and "Michael Tilson Thomas" as a full composer name, generating a false positive. Another example of false positive with homonyms happened with Bernard Haitink, who is a Maestro and had Bernard annotated as the first name of a composer. The same was observed for "York", which is the first name of a composer, and was also annotated as so, in New York. In Silent Woods, Woods was marked as the second name of a composer. This is not a trivial matter.

Another occurrence of false positives linked to names of composers is due to those who have middle names identical to the names of countries. In the biography of Kosaku Yamada, Berlin appears both as a German city and as the composer Berlin Musikhochschule and every time it was marked as the first composer name. Therefore, it generated false positives in cases where it was the city of Berlin. Authors whose names are month names, as April, also generate false positives because in English the names are written with the initially capitalized. However, the author's name will still be annotated properly when it occurs in full, by the first name or middle name, if different from this case. Another interesting and nontrivial case occurs when a composer has one of the names whose term belongs to another class of terms. For example, Ballet. Ballet is a style of composition or musical work, but there is a composer who has Ballet as a second name. In this case, the term receives the two annotations, one of them being false positive. Concert halls with names of composers or personalities that are homonyms of composers also generated false positives.

However, it is interesting to pay attention to the fact that these false positives, has largely been eliminated or received very low weighting in the hierarchical classification process, according the function parameters of relevance calculation, described in Section 4.2, and are not considered to the steps of creating associations and recommending related content. This important factor is clearly perceived to composer names when their first names have generated false positives. As composers are cited in most of the text by their full names or by their middle names, they obtained a much greater relevance factor than their first names, making that the first names stay, for the most part, outside the ontological association process and automatic recommendation, which is the main objective of this work and not the process of annotation.

Still, for the false positives that come to compose associations and entering the recommendation process, in most of the times the recommended documents from them receives, in the final step, a lower score than those that were recommended from true positives. This demonstrates how the proposed methodology acts as expected from the implementation of all of its processes, providing mechanisms to minimize the number of false positives in the related content recommendation. However, these false positives that will eventually be considered yet can impact the end of the process, generating recommendations not relevant and, therefore, the last filter to the recommendations

generated, before the aggregation of the content, is performed manually by the person responsible for its composition.

The use of a single list of full names could be presented as a possible solution to the problem of different markings on names of composers, such as Michael Tilson Thomas, but would create a problem with false negatives, when the authors are referenced only by the second name, which is very common in the literature. The use of two lists in the dictionary, one with full names and one with middle names, in addition to the resolution of homonyms and duplication, might be a plausible solution.

To the problem of composers with names of countries, the use of a trivial solution, such as inserting a list of country names, would seem something plausible at first glance, but it could generate a large number of non-relevant marks, more than the number of the false positives generated by the names of composers. In addition, two annotations would be generated to a composer name, a true positive indicating that this is a composer name, and a false positive indicating that is a country name, plus the annotations in the case when the term is really a country name, having again a true positive and a false positive. Thus, it was concluded that "Countries" is not a class of relevant terms to the music domain.

In the case of homonyms between composers, one possible approach would be to subdivide the composers classification, grouping them by historical periods or music genres and creating up heuristics based on that relationship. Anyway, the solution to namesake is a non-trivial case. To related content recommendation, for example, identify exactly which of the composers of the same name a given text refers may require the analysis of a combination of other elements of the text, to try to associate, for example, the name of the composer to the historical period or to the musical works, to which the text refers. However, assigning a score to recommended documents, as done in this study and described in Section 4.4, presents itself as an efficient resource to reduce the possibility that the content author select, in the end, a misguided text as to the existence of a relationship with the main content, to which is being recommended.

During the analysis of manual annotation process, it was identified that important places, such as opera houses, theaters and renowned music schools were noted recurrently by the collaborators, but are not represented in the ontology. Similarly, it is clear that in the texts that many terms and concepts relevant to the domain are not noted for being in another language, such as French, German and Italian, as is common in music domain. In these cases, have always false negatives for the automatic annotation process, what will incur, in the case of the proposed methodology, in the absence of recommendation for such terms.

The terms that are the same but which are identified by different annotations, do not have their absolute frequencies summed, in the proposed relevance function, being interpreted individually, as they are noted. For example, "Ernesto Lecuona", "Ernesto" and "Lecuona," were annotated correctly in the document about this author, but as independent terms and thereby were weighted. If their weights were added and they appear as a single term in the relevance vector for the document, it would probably associated with the author's name a much higher value relevance. However, it is clear that even recorded separately, the terms appear in the first places in the hierarchical classification by

relevance, being “Lecuona” of relevance equal to 6.0, “Ernesto” with 4.5 and “Ernesto Lecuona” with 3.0. This ensures that the author is considered first in the content recommendation phase.

As can be noticed, there are several factors to be addressed and that can contribute to improve the results obtained. Some of these factors are strongly related to problems of information retrieval, but also point to the possibility of improvements in the organization of the domain knowledge base and the structuring of the ontology. However, the proposed methodology showed to be consistent in the sense of minimizing the impacts of most of these factors on the process of recommending related contents as a whole, producing the expected results when implemented in the prototype of a Recommendation System and Related Content Aggregation, which made possible its execution and empirical verification of the results obtained.

7. Conclusions and Future Work

In this article, we proposed a methodology to automatically recommend related LOs according to SCORM, which enables the relations between LOs by using the relation metadata category, as defined by the standard, and without the need to use or develop a specific LMS to interpret such metadata. In addition, different from other approaches in literature, LOs are recommended as related content to previously selected content of reference, in order to build a learning unit, and are not based on a search process for related content carried by the user.

We developed one domain knowledge base, containing an extensive dictionary of terms and an ontology able to present a conceptualization of the domain area used and the establishment of relations between the various concept classes. This knowledge base was effectively employed by the prototype of the Related Content Recommendation System, also developed in this research to implement the proposed methodology. The following system modules were implemented to fulfill each stage of the methodology: *AssignerRelevance* for automatic metadata annotation and hierarchical classification of relevant concepts; *AssociationsBuilder* for generating associations between the relevant concepts and the domain ontology; *RecommendationsBuilder* for generating related content recommendations, from the pre-established associations; and *DocScoreRecommendationsBuilder*, which gives a final score for each recommended document. The ANNIE plugin was used only for the process of automatic metadata annotation, using the knowledge base.

Based on the analysis of the tests, it was observed that the proposed method in this study is feasible and produces the expected results, with precision and efficiency superior to those achieved only by humans. It can be applied to different areas of knowledge, for the composition of didactic and pedagogical content, requiring only the use of a domain knowledge base related to the desired area, proving the truth of the hypothesis presented earlier. Therefore, the automatic recommendation of related LOs can help e-Learning content developers in their work of LO composition in accordance with SCORM, reducing the time and effort needed to develop and aggregate related content and facilitating reuse.

There are various possibilities of future work and improvements in the results obtained. Due to the fact that the proposed methodology is heavily dependent on the knowledge base, false negatives can occur to terms and concepts that are not present in it. Reducing the number of false positives in the metadata generation and annotation process is of utmost importance since they can impact whole process, generating false positive also in recommended documents. This shows that the accuracy of the recommendation process is related to the accuracy of the annotation and hierarchical classification phase. Possible approaches to reduce these occurrences consist in solving the previously raised problems of information retrieval and also the evaluation of different metrics for the relevance calculation function. The first approach would seek to directly reduce the number of false positives on annotations and the second would seek to reduce the number of annotated false positives transferred to the step of association generation. The identification of new classes in the ontology can expand the coverage of the related content recommendations under the domain area.

Therefore, this research has achieved the proposed objectives, with an effective methodology for the recommendation of related LOs, in accordance with SCORM. The system that implemented the methodology obtained positive results in the processes of automatic metadata annotation, ontological association for the identification of relationships and automatic recommendation of related content. However, there is still much to research and advance in improving these processes such as experimentation and comparison of different metrics to calculate relevance; review, expansion and modification of the domain ontology, considering multiple languages or applying strategies to reduce false positives and resolution of homonyms; proposal of improvements in the methodology.

Acknowledgment

We are grateful to the Brazilian funding agencies FAPEMIG (grants 2013 and 2014) and CAPES for their financial support.

References

Advanced Distributed Learning. (1999). *The Advanced Distributed Learning (ADL) Initiative*: history. `http://www.adlnet.gov/overview`

Advanced Distributed Learning. (2009a). *SCORM 2004: content aggregation model [CAM] (4th ed.)*. `http://www.adlnet.gov/scorm/ scorm-2004-4th/`

Advanced Distributed Learning. (2009b)**.** *SCORM 2004: run-time environment [RTE] (4th ed.)*. `http://www.adlnet.gov/scorm/scorm-2004-4th/`

Advanced Distributed Learning. (2009c). *SCORM 2004: sequencing and navigation [SN] (4th ed.)*. `http://www.adlnet.gov/scorm/ scorm-2004-4th/`

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., Aswani, N., Roberts, I. *et al.* (2012). *Developing Language Processing Components with GATE Version 7 (a User Guide)*. Sheffield: University of Sheffield. `https://gate.ac.uk/sale/tao/tao.pdf`

Dorça, F.A., Carvalho, V.C. de, Resende, D.T., Cattelan, R.G. (2016). An automatic and dynamic approach for personalized recommendation of learning objects considering students learning styles: an experimental analysis. *Informatics in Education*, 15(1), 45–62.

Edvardsen, L.F.H., Intell. Commun. AS, Oslo, N., Solvberg, I.T., Aalberg, T., Tratteberg, H. (2009). Using automatic metadata generation to reduce the knowledge and time requirements for making SCORM learning objects. In: *Proceedings of the 3rd IEEE International Conference on Digital Ecosystems and Technologies*. Istanbul, Turkey, 253–258.

Engelhardt, M., Hildebrand, A., Dagmar L., Schmidt T. C. (2006). Reasoning about e-learning multimedia objects**.** In: *Proceedings of the 1st International Workshop on Semantic Web Annotations for Multimedia*. Edinburgh, Scotland.
http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.120.8484&rank=1

Guo, Z., Jin, H. (2011a). A rule-based framework of metadata extraction from scientific papers. In: *Proceedings of the 10th International Symposium on Distributed Computing and Applications to Business, Engineering and Science*. Wuxi, China, 400–404.

Guo, Z., Jin, H. (2011b). Reference metadata extraction from scientific papers. In: *Proceeding of the 12th International Conference on Parallel and Distributed Computing, Applications and Technologies.* Gwangju, South Korea, 45–49.

Hernández, A., Badell, C., Sum, R., Motz, R. (2009). Convirtiendo el contenido de archivos en objetos de aprendizaje. In: *Proceedings of the 20th Brazilian Symposium of Computer in Education.* Florianópolis, Brazil.
http://www.niee.ufrgs. br/eventos/SBIE/2009/conteudo/artigos/completos/62173_1.pdf

Huynh T., Hoang, K. (2010). GATE framework based metadata extraction from scientific papers. *In Proceedings of ICEMT10 – International Conference on Education and Management Technology*. Cairo, Egypt, 188–197.

Lipinski, M., Yao, K., Breitinger, C., Beel, J., Gipp, B. (2013). Evaluation of header metadata extraction approaches and tools for scientific PDF documents. In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*. New York, United States of America, 385–386.

Lu, E. J.-L., Hsieh, C.-J. (2009). A relation metadata extension for SCORM content aggregation model. *Computer Standards & Interfaces*, 31(5), 1028–1035.

Lu, E. J.-L., Horng, G., Yu, C.-S., Chou, L.Y. (2010). Extended Relation Metadata for SCORM-based Learning Content Management Systems. *Educational Technology & Society*, 13(1), 220–235.

Maratea, A., Petrosino, A., Manzo, M. (2012). Automatic generation of SCORM compliant metadata for portable document format files. In: *Proceedings of the 13th International Conference on Computer Systems and Technologies*. Ruse, Bulgaria, 360–367.

Margaritopolous, M., Manitsaris, A., Mavridis, I. (2007). On the identification of inference rules for automatic metadata generation.
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.108.1723&rep=rep1&type=pdf

Maynard, D. (2008). Benchmarking Textual Annotation Tools for the Semantic Web. In: *Proceedings the 6th International Conference on Language Resources and Evaluation*. Marrakech, Morocco.
http://www.lrec-conf.org/proceedings/lrec2008/

Morais, E.A.M., Ambrósio, A.P.L. (2007). *Mineração de textos*. Technical Report. Goiânia, Brazil. Federal University of Goiás. Access: January 22, 2014. Retrieved from:
http://www.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_005-07.pdf

Nauerz, A., Bakalov, F., König-Ries, B., Welsh, M. (2008). Personalized recommendation of related content based on automatic metadata extraction. In: *Proceedings of the 2008 Conference of the Center for Advanced Studies on Collaborative Research: Meeting of Minds*. Ontario, Canada.

Rey-López, M., Díaz-Redondo, R.P., Fernández-Vilas, A., Pazos-Arias, J.J., García-Duque, J., Gil-Solla, A. *et al.* (2009). An extension to the ADL SCORM standard to support adaptivity: The t-learning case-study. *Computer Standards & Interfaces*, 31(2), 309–318.

Roy, D., Sudeshna S., Sujoy, G. (2008). Automatic extraction of pedagogic metadata from learning content. *International Journal of Artificial Intelligence in Education*, 18(2), 97–118.

Solomou, G., Pierrakeas, C., Kameas, A. (2015). Characterization of educational resources in e-learning systems using an educational metadata profile. *Educational Technology & Society*, 18(4), 246–260.

Su, J.-M., Tseng, S.-S., Chen, C.-Y., Weng, J.-F., Tsai, W.-N. (2006). Constructing SCORM compliant course based on high-level petri nets. *Computer Standards & Interfaces*, 28(3), 336–355.

Tuarob, S., Pouvhard, L.C., Giles, C.L. (2013). Automatic tag recommendation for metadata annotation using probabilistic topic modeling. In: *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital libraries*. New York, United States of America, 239–248.

Tarus, J.K., Niu, Z., Yousif, A. (2017). A hybrid knowledge-based recommender system for e-learning based on ontology and sequential pattern mining. *Future Generation Computer Systems*, 72, 37–48.

D.E. Neves is a PhD student in Computer Science from the Postgraduate Program in Computing at the Pontifical Catholic University of Minas Gerais (PUC Minas), with concentration in Computer Science and acting in the line of research in Data Analysis, Knowledge Discovery and Information Retrieval. He has a Masters in Informatics, Specialization in Development of Applications for WEB and Technological Graduation in Digital Games, all three from PUC Minas. With a multidisciplinary background, he is a teacher with a Full Degree in Performing Arts and a Bachelor in Performing Arts, both from the Fine Arts School of the Federal University of Minas Gerais. He has been working with software development since 2008, working on several projects, including analysis, architecture and systems development. He also has experience in developing solutions for e-Learning and educational games. In the area of education, he worked as a teacher since the completion of the Bachelor in 2003, teaching arts for elementary education, high school and youth and adult education until 2010.

W.C. Brandão is associate professor in the Department of Computer Science at Pontifical Catholic University of Minas Gerais (PUC Minas) in Belo Horizonte, Brazil. Research interests include information retrieval, database systems, data mining, knowledge discovery, machine learning, and social network analysis.

L. Ishitani is a computer science Professor at Pontifical Catholic University of Minas Gerais (PUC Minas), Belo Horizonte, Brazil. She has a BSc (1990), a MSc (1993) and DSc (2003) in Computer Science from Federal University of Minas Gerais (UFMG), Belo Horizonte, Brazil. Her research interests are focused on computers in education, learning objects, serious games and interaction design. She works as a reviewer for several journals and conferences in the field of informatics in education.